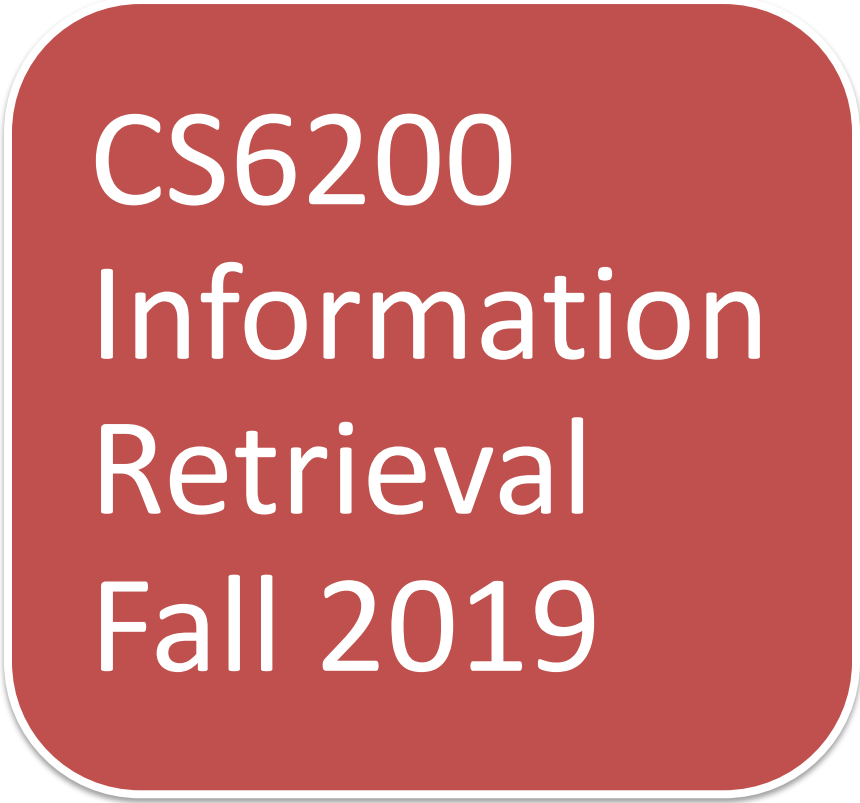


Northeastern University - Seattle

Khoury College of  
Computer Sciences

Lecture 1: Introduction  
to Information Retrieval

Sep 9, 2019

A red rounded rectangle with a white border, containing the course title and semester in white text.

CS6200  
Information  
Retrieval  
Fall 2019

# Overview

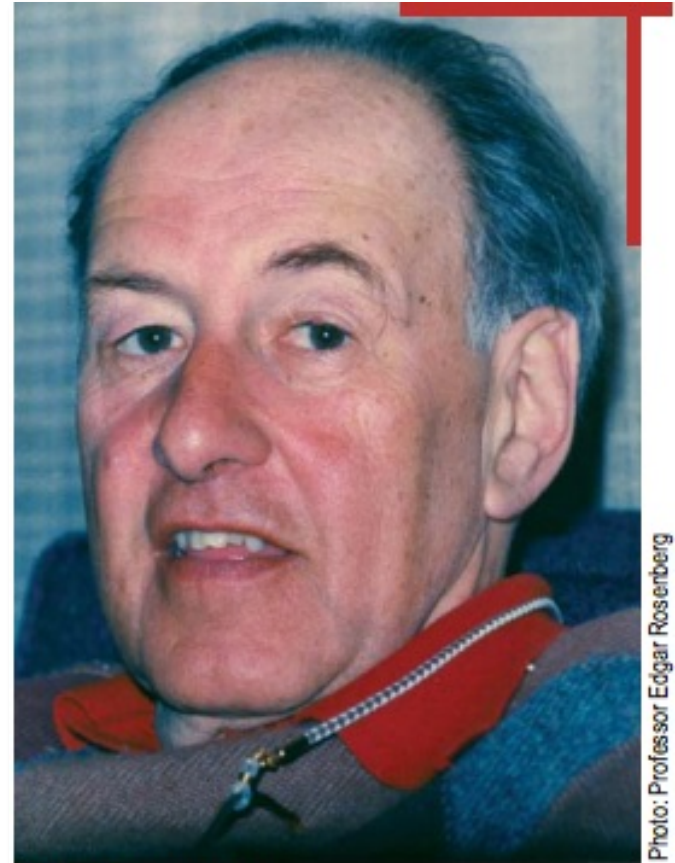
- Definitions of Information Retrieval (IR)
- IR Systems
- Databases
- What makes IR Systems different
- (Web) Search Engines
- What makes Search Engines different
- Why is Search important

# DEFINITIONS OF IR

# Definitions of IR

“Information Retrieval (IR) is concerned with the representation, storage, organization, and accessing of information items”

Gerard Salton/  
Michael McGill, 1983



Gerry Salton, 1927-1995

Picture from  
[http://www.cs.cornell.edu/qries/40brochure/pg24\\_25.pdf](http://www.cs.cornell.edu/qries/40brochure/pg24_25.pdf)

# Definitions of IR

“Information Retrieval (IR) is finding material (usually documents), of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)”

Manning, Raghavan and Schütze, 2008

# Digging into the Definition

- IR all around us
- Definitions intuitively clear
- But what are:
  - Documents? Nature of documents?
  - Information needs?
  - Collections?
- And how does IR compare with Database systems or Web search?

# IR SYSTEMS

# Documents & The Nature of Documents



Varying sizes: As small as a tweet, as big as a book, or even as huge as an encyclopedia



Various types: web pages, news stories, books, papers, blogs, IMs, tweets, Facebook posts, PDFs, Office documents, etc. etc.



May have definite structure:

mail messages (From/To/Subject);  
papers (Authors, Affiliations, Where published...)  
product details (as on shopping sites)



Typically content is (mostly) text, but may include images, audio and video clips etc.



# Information Needs



A specification of information the user is looking for. Examples:

“Picture of Gerard Salton”

“Women’s Singles winner in the US Open this year?”

“What’s a good vegetarian place we could go to, tonight?”

“Next total solar eclipse and where I can see it”



Need to understand info need and retrieve appropriate information



Information needs get mapped to queries

# Collections



Could be very general, e.g. web documents for Bing/Google/Baidu/Yandex etc. (“Internet”)



Could be specific to organizations (“Intranet”):

All NEU documents  
All Microsoft documents



Could be domain-specific:

All scientific papers  
(DBLP, ArXiv etc.)



Could be small and specific:

Enron email corpus

# DATABASES

# Databases



Databases could be large but mostly “about” a well-defined domain



Typically very structured



Usually clearly defined tables, records and relations

Examples:  
Company and employee tables  
Product info tables, with records for each product  
...

# Database queries & results ...1



Typically very specific

E.g. “Find employees with salary > \$100k in Seattle”



Easy to understand query

Semantics of query language usually unambiguous  
Typically only one way to interpret query



Results usually very clear

Because query is clear, results equally clear

# Database queries & results ...2



Ordering of results usually  
well-defined  
(or not important)

Usually specified, or standard default, e.g.  
“Find employees with salary > \$100k in Seattle  
order by lastname asc”

Usually all matching records required



Evaluation of results usually simple

# WHAT MAKES IR DIFFERENT

# What makes IR different?

## Not just text



Dealing not just with numbers/text, but with:

Images  
Audio  
Video  
Mail messages  
...



Query interpretation much more than exact matches on text, or numeric comparisons



# What makes IR different? Specificity

Queries  
may be very  
vague:

- E.g. “North Korea missiles”  
What does the user expect?

Boolean  
queries may  
help a bit

- E.g. “North Korea” AND “missiles” NOT ... OR ...
- But hard to construct, and still not enough.  
Is one mention of  
“North Korea” and “missiles” enough to make it “relevant”?  
That is, a good result? → Leads us to “Aboutness”

# What makes IR different? “Aboutness”



User’s query could be hard to interpret:

Many ways to say the same thing

“Who won the US Open Women’s Finals?”,  
“Who is the US Open Women’s Finals winner?”

The same word could mean different things:  
“Tesla”, “Jaguars” vs. “Jaguar”, ...

Does one mention of “Northeastern”  
make it “about” Northeastern things? NEU?



Hard to choose results, and order of results



“Aboutness” related to context and relevance

# What makes IR different?

## Context

- Unlike with databases, **context** is important
  - You and I know different things and expect different results based on our understanding
  - We expect different results based on our location, previous queries etc.

For example:

“Pizza place”

“income tax office”

➔ Evaluation of search is user-centric

# What makes IR different? Relevance



Results have to be “relevant”

A relevant result document is one that the “user perceives as containing information of value with respect to their personal information need” [MKS].



Users typically want results in order of relevance.



Results and their ranking not computed on a complete “understanding” of the document collection and users’ queries, but based on retrieval models

e.g. vector space model, probabilistic model etc.

# What makes IR different?

## Evaluation

- Unlike databases, evaluation of results is non-trivial
  - Many methods to compare search engine results with users' notion of relevant results and result order
    - Users' expectations may even change with time
  - Some methods based on test collections and gold standards
  - Others based on side-by-side (Alpha-Beta) comparisons
  - Many metrics of the goodness (or effectiveness) of retrieval
    - including Precision, Recall, MAP, MRR, DCG, NDCG etc.

# SEARCH ENGINES AND HOW THEY ARE DIFFERENT

# Search Engines

Search engines: IR techniques applied to very large document collections

- Web Search engines: Search engines on web-scale document collections
- Many varieties, based on domain or application

Includes:

- Web search engines: Bing, Google, Yandex, Baidu etc.
- Open source engines: Lucene/Solr, ElasticSearch etc.
- Academic search engines: Lemur, Indri, Terrier etc.
- ...

# What makes (Web) Search Engines different?



Like all big systems, performance and scalability



Dealing with changing data



Different retrieval techniques for different domains



Issues with large-scale Spam and Duplicates and ranking issues



The need to provide a great user experience



Special issues because web search engines are 'free'



# Search Engines: Performance

Performance important, to keep and grow user base

Users have high expectations

- Need to be fast, keep response time low

Behind the scene efficiency requirements:

- Process documents and index them faster
- High query throughput, to handle more users/queries
- High accuracy, precision, recall etc.
- Fast spelling correction, fast rendering, fast suggestions, fast federation etc.

# Search Engines: Scalability

As the collection size grows larger, and as the number of simultaneous users grows bigger,

- Need to keep up, be as accurate and performant

Handling more data and more users:

- More systems? More memory etc.? Better hardware?
- Improved processes?
- Different architecture: e.g. more distributed systems?
- Improved software architecture?

# Search Engines:

## Keeping up with data

Data collections keep growing at an impressive rate

- Especially Social media data (Twitter, Facebook, Weibo...)

Not all data updated frequently, or at the same rate

But users expect

- Coverage: more documents to be indexed and served
- Freshness: latest versions of documents

Need to get data as often as it is refreshed

Need to index and update running systems

# Search Engines: Different domains



Different domains need different strategies and techniques for data acquisition, indexing, retrieval, presentation...

Consider web documents vs. news vs. product search vs. image or video search



Need to handle all this with a consistent UI/UX

# Search Engines: Spam/Duplicates

- The web has loads of
  - Duplicate documents
    - E.g. GNU License document
  - Spam
    - Could come from SEO tools (term or link spam)
- Need to identify and deal with duplicates and spam, to improve performance and efficiency

# Search Engines: User Experience

Need to give users a great search experience

- To impress and help users
- To keep users from going to other engines

Need to balance features against cost

# Search Engines: “Free” search

Need to cover costs

- Need for ads, sponsored sites etc.

Again, need to consider user experience

- Good ads can complement web results
- But too many ads may chase away users

# Our Focus

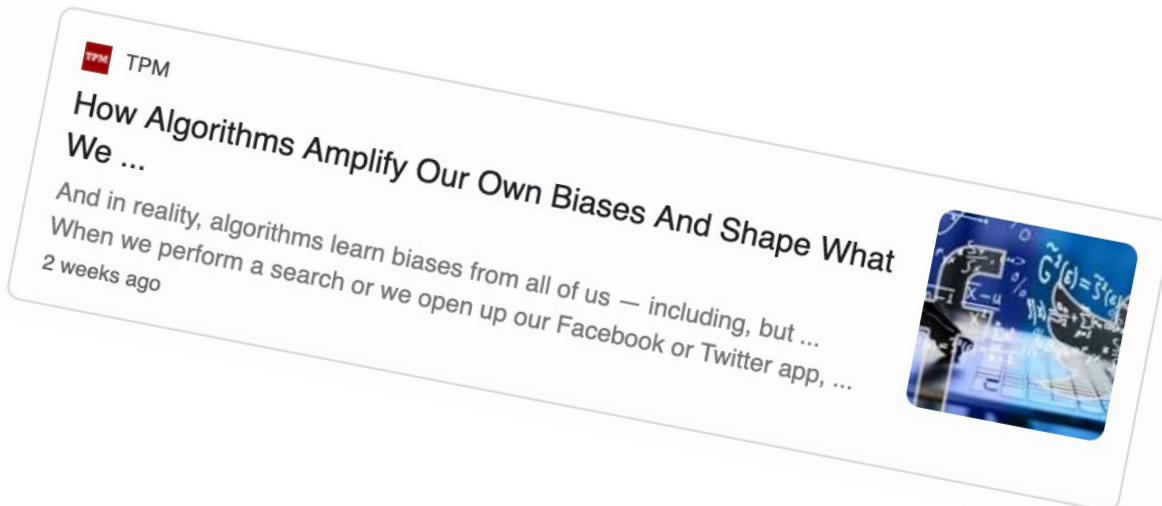
- We will focus on (Web) Search Engines, and:
  - Study individual search engine components, and techniques and alternatives used in these
  - Emphasize architectures and techniques used in real search engines
  - Provide pointers to follow up on other approaches



# WHY IS SEARCH IMPORTANT

# Why is Web Search so important? .. 1

- Results can show bias in civic discourse
- Should offer access to all voices (?)



## Dr. Robert Epstein: Study claims Google reflected 'very dramatic bias' in 2016 election search results

Google allegedly offered search results during the 2016 election season that manipulated voters in Hillary Clinton's favor, ...

Fox News | 16h



# Why is Web Search so important? .. 2

## Marketing 'bias'

amazon prime

Deliver to Raman  
Seattle 98103

Your Pickup Location    [Browsing History](#)    [Today's Deals](#)    [Raman's Amazon.com](#)    [Buy Again](#)    [Gift Cards](#)    [Help](#)    [Credit Cards](#)    [Whole Foods](#)    [Amazon Business](#)    [Sell](#)

All Electronics    [Deals](#)    [Best Sellers](#)    [TV & Video](#)    [Audio & Home Theater](#)    [Computers](#)    [Camera & Photo](#)    [Wearable Technology](#)    [Car Electronics & GPS](#)    [Portable Audio](#)    [Cell Phones](#)    [Office Electronics](#)    [Musical Instruments](#)    [New Arrivals](#)    [Trade-In](#)

1-24 of over 20,000 results for **Electronics: "hdmi cable"**

**Amazon Prime**  
☒ prime

**Department**  
Any Department  
**Electronics**  
HDMI Cables  
Computer Cables & Interconnects  
Fiber Optic Cables  
Audio & Video Accessories  
Electronics Warranties  
[See more](#)

**Avg. Customer Review**  
★★★★☆ & Up  
★★★★☆ & Up  
★★★★☆ & Up  
★★★★☆ & Up

**Brand**  
☐ AmazonBasics  
☐ BlueRigger  
☐ Monster  
☐ Mediabridge  
☐ SecurOMax  
☐ C&E  
☐ Aurum Cables  
☐ Twisted Veins  
☐ IBRA  
☐ Syncwire  
☐ KabelDirekt  
☐ Cablelera  
☐ Rankie  
☐ Cable Matters  
[See more](#)

**Price**  
Under \$25  
\$25 to \$50  
\$50 to \$100  
\$100 to \$200  
\$200 & Above  
\$Min \$Max

**Cable Length**  
☐ Under 4 Feet  
☐ 4 to 5.9 Feet  
☐ 6 to 7.9 Feet  
☐ 8 to 9.9 Feet  
☐ 10 to 14.9 Feet  
☐ 15 to 24.9 Feet  
☐ 25 Feet & Above

**International Shipping**  
☐ International Shipping Eligible

**Cable Connection Gender**  
☐ Male-Male  
☐ Female-Male

**Cable Color**  
☒ Black  
☐ Blue  
☐ Red  
☐ White

**Cable & Interconnect Length**  
☐ Under 4 Feet  
☐ 4 to 5.9 Feet  
☐ 6 to 7.9 Feet  
☐ 8 to 9.9 Feet  
☐ 10 to 14.9 Feet  
☐ 15 to 24.9 Feet  
☐ 25 Feet & Above

**Electronic Warranty Length**  
☐ 1 Year

**SPONSORED BY ATEVON**  
Ultra High Speed HDMI 2.0 Cable for Video, Games  
[Shop now](#)

4K HDMI Cable 6 ft - Atevon  
★★★★☆ 2,802  
prime

4K HDMI Cable 10 FT - Atevon  
★★★★☆ 2,802  
prime

4K HDMI Cable 3.3 Feet - Atevon  
★★★★☆ 2,802  
prime

**Sponsored**  
AmazonBasics Braided 4k HDMI to HDMI Cable - 10-Foot  
★★★★☆ 482  
**Electronics**  
\$8.95 \$12.99  
prime FREE Delivery Tue, Sep 10

**Sponsored**  
4K HDMI Cable 6.6 ft, iVanky High Speed 18Gbps HDMI 2.0 Cable, 4K HDR, 3D, 2160P, 1080P, Ethernet - Braided HDMI Cord 30AWG, Audio Return (ARCT) Compatible UHD TV, Blu-ray, Xbox, PS4, PC, BT  
★★★★☆ 469  
**Electronics**  
\$9.99  
prime FREE Delivery Tue, Sep 10

**Top rated from our brands**  
Amazon's private and select exclusive brands.  
[See more](#)

**Amazon's Choice**

AmazonBasics High-Speed 4K HDMI Cable, 6 Feet, 1-Pack  
★★★★☆ 21,089  
**Electronics**  
\$6.99  
prime

AmazonBasics Braided 4k HDMI to HDMI Cable - 10-Foot  
★★★★☆ 482  
**Electronics**  
\$8.95 \$12.99  
prime

AmazonBasics CL3 Rated High Speed 4K HDMI Cable - 15 Feet  
★★★★☆ 1,441  
**Electronics**  
\$10.99  
prime

AmazonBasics DVI to HDMI Cable  
★★★★☆ 4,567  
**Electronics**  
\$7.99  
prime

AmazonBasics High-Speed Male to Female HDMI Extension Cable - 6 Feet  
★★★★☆ 1,060  
**Electronics**  
\$6.99  
prime

**Amazon's Choice**

AmazonBasics High-Speed 4K HDMI Cable, 6 Feet, 1-Pack  
★★★★☆ 21,089

# Why is Web Search so important? .. 3



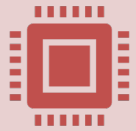
Search is how we navigate through the world

Definitions

Translations

News

Directions ...



Complex software, involving big data, machine learning, text analysis, ranking, operations: a big exercise in software project management



Affects everything.  
Important to study it & get it **right**

# Summary

- We've defined Information Retrieval (IR) and IR systems
- We've seen how they're different from Databases
- We've then seen how Search Engines differ from generic IR systems
- We've defined our focus: (Web) Search Engines, components, architecture and techniques
- Next: we look at ***Search Engine Architecture***

# Readings

- Chapter 1 of CMS
- Recent news articles (just a sample):
  - <https://www.foxnews.com/media/google-bias-search-results-trump-clinton-epstein-levin>
  - <https://psychcentral.com/blog/dr-epstein-political-bias-google-search-results/>

# Questions?